CLIMB, OpenStack & ~~GPFS~~ Elastic Storage

Simon Thompson
Research Computing Team
University of Birmingham, England, UK

UNIVERSITY OF BIRMINGHAM

University of Birmingham

- Research intensive University
- ~19000 Undergraduate Students
- ~6400 Postgraduate Taught
- ~2900 Postgraduate Research
- £145.5 million (~$230 million) in research income (2011-12)

Data for 2011/2012 academic session

CLIMB Project

- Funded by Medical Research Council (MRC)
- Four partner Universities
  - Birmingham
  - Cardiff
  - Swansea
  - Warwick
- ~£8m (~$13M) grant
- Private cloud, running 1000 VMs over 4 sites

UNIVERSITY OF BIRMINGHAM

The CLIMB Consortium

- Professor Mark Pallen (Warwick) and Dr Sam Sheppard (Swansea) – Joint PIs
- Professor Mark Achtman (Warwick), Professor Steve Busby FRS (Birmingham), **Dr Tom Connor (Cardiff)***, Professor Tim Walsh (Cardiff), Dr Robin Howe (Public Health Wales) – Co-Is
- **Dr Nick Loman (Birmingham)*** and Dr Chris Quince (Warwick) ; MRC Research Fellows

** Principal bioinformaticians architecting and designing the system*

UNIVERSITY OF BIRMINGHAM

The CLIMB Consortium

- Professor Mark Pallen (Warwick) and Dr Sam Sheppard (Swansea) – Joint PIs
- Professor Mark Achtman (Warwick), Professor Steve Busby FRS (Birmingham), **Dr Tom Connor (Cardiff)***, Professor Tim Walsh (Cardiff), Dr Robin Howe (Public Health Wales) – Co-Is
- **Dr Nick Loman (Birmingham)*** and Dr Chris Quince (Warwick) ; MRC Research Fellows

** Principal bioinformaticians architecting and designing the system*

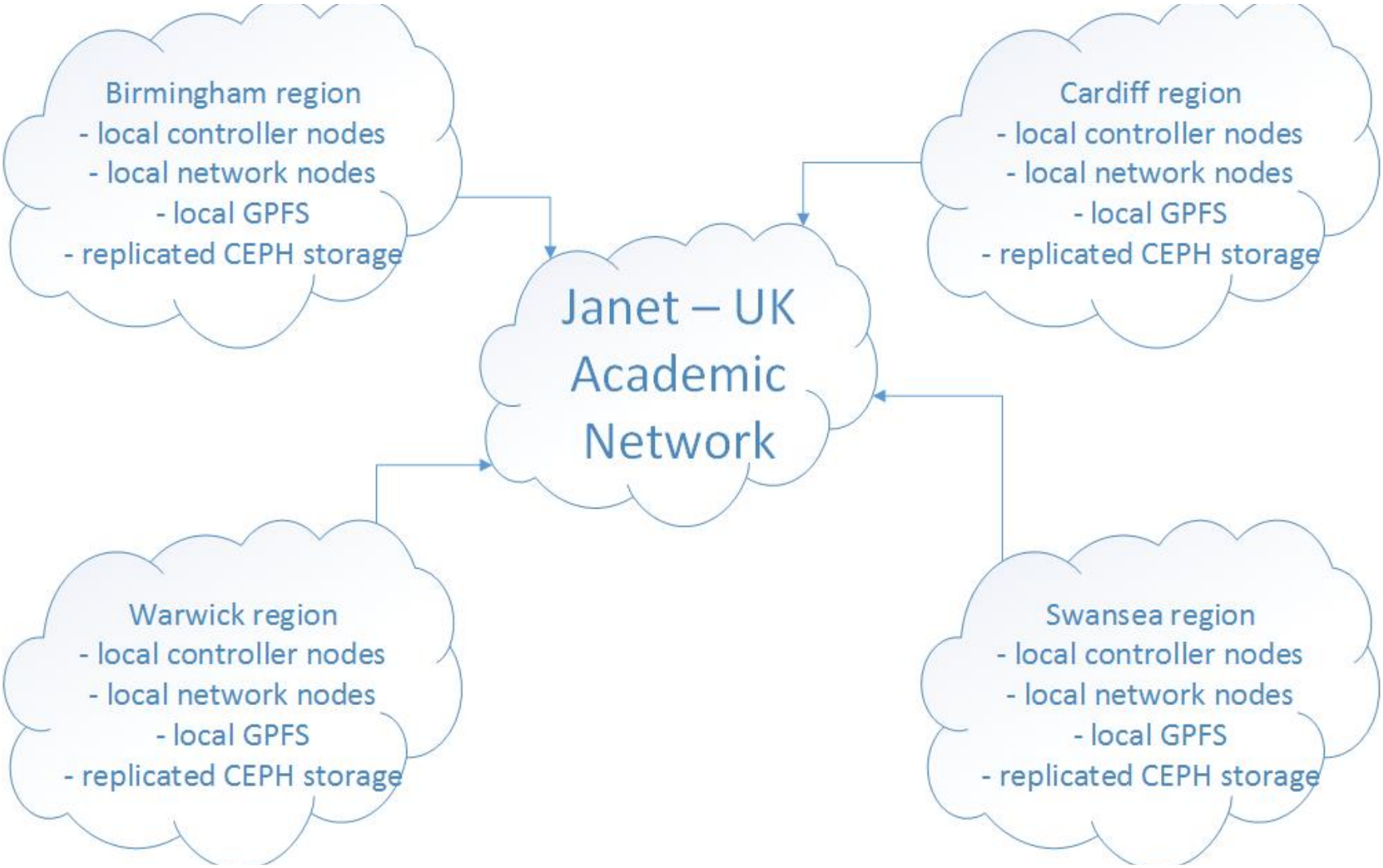
UNIVERSITY OF BIRMINGHAM

CLIMB

- Separate OpenStack region per site
- Federated single gateway to access
- Local GPFS high performance
  - ~0.5PB per site
- CEPH storage cluster replicated across sites
  - For archive of VMs
  - Between 2-5PB total usable over 4 sites

UNIVERSITY OF BIRMINGHAM

CLIMB Overview



Our stack

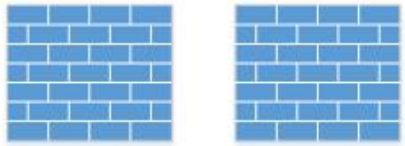
- GPFS 4.1.0 PTF 3
- Scientific Linux 6.5
- Openstack Icehouse release (Redhat RDO)

- Mellanox FDR-14 IB
- Mellanox 40GbE Ethernet
- IBM (Lenovo?!) System X hardware
- Brocade VDX 10/40GbE switches

UNIVERSITY OF BIRMINGHAM

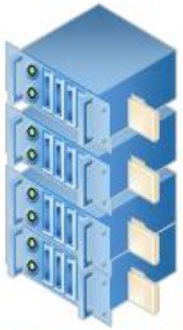
Public Network

HA Firewall Pair
Brocade Vyatta v5650 vrouter
IBM x3650m5 (Haswell)



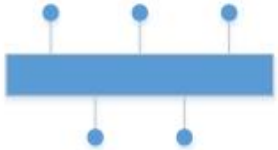
Site to site replicate over VPN Tunnel

OpenStack controller nodes
Neutron network nodes
IBM x3650m4



CEPH Storage (Dell)
Replicated locally and at least 1 remote copy (spec & config TBC)

Mellanox SX6036 FDR-14
Mellanox SX1036 40GbE
Ethernet switch
Brocade VDX

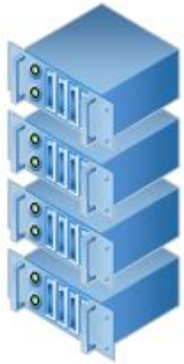


Infiniband FDR-14 for GPFS & bare metal MPI jobs?
40GbE for GPFS, OpenStack management, Openstack private
1GbE management (xcat)
10GbE (20GbE bond) CEPH, OpenStack private

GPFS Servers
IBM x3650m4

Direct SAS multipath

V3700 Storage Arrays
9x 24 disk 4TB expansions



OpenStack (nova compute) nodes
IBM x3750m4 (512GB RAM)
IBM x3960 x6 (3TB RAM)

Block size alignment

- V3700 – 256KB strip by default
- = 2MB stripes (8+2P raid sets)
- Data sets are likely to be large (100GB?), or storing VM images
- 4KB inodes allow small files to be stored in inode
- 8MB blocks

UNIVERSITY OF BIRMINGHAM

File-system considerations

- Pre-allocate large number of inodes
- pagepool – 30-50% of node memory
- maxFilesToCache
- maxStatCache (4x maxFilesToCache)
- seqDiscardThreshold

GPFS magic sauce & OpenStack

- Swift
  - Object storage
- Glance
  - Image service (where we store VM images)
- Cinder
  - Volume (block disk service)
- Nova compute
  - The bit that runs on the Hypervisor servers

UNIVERSITY OF BIRMINGHAM

Swift (object storage)

- Runs directly on GPFS servers
- Clients connect to swift via API
- Shared file-system so no need to replicate objects between glance nodes
- Use a separate file-set for swift

UNIVERSITY OF BIRMINGHAM

Swift

- There’s an IBM red paper on it
  - Set object replication at 1 (GPFS provides access and replication if needed)
  - Set replication factor of account/container rings at 2 or 3
  - 10 vdevices per swift node
  - Pre-allocate inodes for performance (we have 200M inodes allocated)
  - Don’t use GPFS ACLs or Quotas

UNIVERSITY OF BIRMINGHAM

Glance (image service)

- Share file-set with Cinder
- Set in both glance-api and glance-cache:
- filesystem_store_datadir = /climb/openstack-data/store
- default_store = file
- Ensure you have glance.store.filesystem.Store in known_stores
- Ensure that the directory is writable!

UNIVERSITY OF BIRMINGHAM

Cinder (Volume service)

- GPFS driver for Cinder in OpenStack
- Allows glance image provision by GPFS snapshot
- Copy on write

UNIVERSITY OF BIRMINGHAM

Nova compute

- Point Nova compute at GPFS
- It’s a shared file-system so can live migrate
  - Horizon confused about space
- Normal GPFS storage so can use RDMA
- Will LROC improve performance here?

UNIVERSITY OF BIRMINGHAM

What would be nice?

- Direct access to GPFS file-system from VMs
  - VirtIO with KVM? OpenStack support?
  - GPFS client? … but how would it network
  - UID mapping?

Future GPFS work

- Tune GPFS environment – any thoughts?
- Add local SSDs to enable LROC for nova-compute nodes?
- AFM to replicate glance across sites
- Integrate OpenStack environment with GPFS and CEPH storage

UNIVERSITY OF BIRMINGHAM

GPFS @UoB

- BlueBEAR – Linux HPC running over FDR-10
- Research Data Store – multi-data centre, replicated, HA failover system for bulk data for research projects
- Hadoop?

UNIVERSITY OF BIRMINGHAM

More Info/Contact

- Me: S.J.Thompson@bham.ac.uk
- www.roamingzebra.co.uk (shameless blog plug)
- Project: www.climb.ac.uk
- Twitter: @MRCClimb

UNIVERSITY OF BIRMINGHAM